How Aligned are Human Chart Takeaways and LLM Predictions? A Case Study on Bar Charts with Varying Layouts

Huichen Will Wang, Jane Hoffswell, Sao Myat Thazin Thane, Victor S. Bursztyn, and Cindy Xiong Bearfield

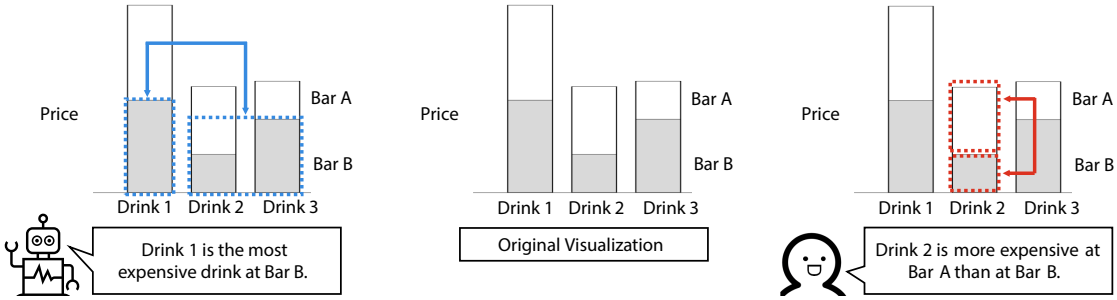

Fig. 1: There is a discrepancy between human chart takeaways and predictions of human chart takeaways generated by large language models. For a chart that shows the price of three drinks in two bars, a human would tend to compare the price of Drink 2 between the two bars, but the model predicts a human to compare the price of the three drinks in Bar B.

Abstract—Large Language Models (LLMs) have been adopted for a variety of visualizations tasks, but how far are we from perceptually aware LLMs that can predict human takeaways? Graphical perception literature has shown that human chart takeaways are sensitive to visualization design choices, such as spatial layouts. In this work, we examine the extent to which LLMs exhibit such sensitivity when generating takeaways, using bar charts with varying spatial layouts as a case study. We conducted three experiments and tested four common bar chart layouts: vertically juxtaposed, horizontally juxtaposed, overlaid, and stacked. In Experiment 1, we identified the optimal configurations to generate meaningful chart takeaways by testing four LLMs, two temperature settings, nine chart specifications, and two prompting strategies. We found that even state-of-the-art LLMs struggled to generate semantically diverse and factually accurate takeaways. In Experiment 2, we used the optimal configurations to generate 30 chart takeaways each for eight visualizations across four layouts and two datasets in both zero-shot and one-shot settings. Compared to human takeaways, we found that the takeaways LLMs generated often did not match the types of comparisons made by humans. In Experiment 3, we examined the effect of chart context and data on LLM takeaways. We found that LLMs, unlike humans, exhibited variation in takeaway comparison types for different bar charts using the same bar layout. Overall, our case study evaluates the ability of LLMs to emulate human interpretations of data and points to challenges and opportunities in using LLMs to predict human chart takeaways.

Index Terms—Visualization, Graphical Perception, Large Language Models

1 Introduction

Designing a visualization involves a series of decisions, ranging from choosing the chart type and selecting a color palette, to determining the amount of accompanying text. The data visualization literature has shown that these design choices can significantly influence how readers perceive and interpret the data presented. For instance, when presenting the same data, visualizations with higher levels of data aggregation, like bar charts, often lead readers to perceive stronger causal relationships than visualizations with lower levels of aggregation, such as line charts and scatterplots [40]. While bar charts more readily afford discrete comparisons between data points, line charts are better suited for identifying trends [44]. Moreover, choosing a suitable visualization format can even help mitigate confirmation bias [38].

Even when the same chart type is used to visualize a dataset, viewer perception and takeaways may diverge based on other design choices. For bar charts, factors like ordering, partitioning, and spacing are all tied to affordances of different message types [14]. Different spatial layouts of bars also afford different types of comparisons [3, 15]. For instance, when bars are overlaid (Figure 2), readers are more likely to identify the maximum and minimum values. However, when bars are adjacently aligned, readers become more likely to pick one bar and compare it to multiple other values. The richness of design choices and their intricate relations to perceptual and cognitive affordances make designing effective visualizations challenging. Even visualization experts can fail to match visualization designs to affordance types in a highly constrained design space of four bar layout possibilities [39].

Large language models (LLMs) have recently taken the world by storm due to their remarkable ability to generate coherent text and follow instructions. Trained on vast text corpora, LLMs have the potential to encode rich visualization knowledge. Thus, they offer exciting new possibilities for visualization tasks. For instance, Chen et al. [11] utilized Codex [10] to transform natural language queries into structurally parameterized SQL, which was used to generate interactive visualization interfaces. More powerful LLMs further simplify visualization generation. For example, GPT-4 allows users to specify chart requirements and directly produces high-quality visualizations in many instances [9].

So can LLMs effectively leverage its knowledge base to account for visualization design affordances? In this paper, we compare the sensitivity of LLMs to changes in visualization design to that of humans' when generating visualization takeaways. Specifically, we explore whether LLMs are perceptually aware of design manipulations like humans are by examining bar charts as a case study. Existing work

- *Huichen Will Wang is with the University of Washington.*
- *Jane Hoffswell and Victor S. Bursztyn are with Adobe Research.*
- *Sao Myat Thazin Thane is with UMass Amherst.*
- *Cindy Xiong Bearfield is with the Georgia Institute of Technology.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

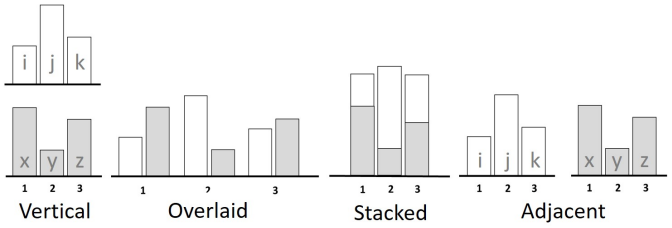

Fig. 2: Figure from Xiong et al. [39] showing four spatial arrangements.

by Xiong et al. [39] has examined in detail how common layouts affect the type of comparisons people make when reading bar charts. The researchers generated two-by-three bar charts using two datasets and visualized them in four layouts (vertically juxtaposed, horizontally juxtaposed, overlaid, and stacked, as shown in Figure 2). They collected a corpus of natural language takeaways via crowdsourcing, classifying the types of comparisons human readers made upon seeing these bar charts into 12 categories. Depending on the spatial layout of the bar chart, a human reader can be more or less likely to make a comparison of a specific type. For example, for a dataset depicting the revenue of three stores (1, 2, and 3) in two companies (A and B), people were more likely to compare the revenue of Company A's Store 1 to that of Company B's Store 1 when reading an adjacent bar chart, as shown in Figure 4, but they do not make that comparison when they read an overlaid or stacked version of the bar chart.

We leveraged the experimental results from Xiong et al. [39] by asking state-of-the-art LLMs to generate natural language takeaways to the same set of bar charts. Through three experiments, we compared the distributions of human takeaways from [39] across the 12 categories to the distributions of LLM takeaways to investigate whether LLMs are perceptually sensitive to changes in visualization designs.

Experiment 1: Identifying the optimal configurations. We experimented with different configurations for generating chart takeaways across four LLMs, two temperatures, nine chart specifications, and two prompting strategies in zero-shot and one-shot settings. We identified representing charts with ggplot2, utilizing GPT-4 with a temperature of 0, and employing the guided discovery strategy as the optimal configuration for the zero-shot setting and representing charts with Vega-Lite and image, using GPT-4V with a temperature of 0.7, and employing the baseline strategy as the optimal configuration for the one-shot setting. In addition, we found that state-of-the-art LLMs could struggle to generate semantically diverse and factually accurate takeaways.

Experiment 2: Human-alignment of LLM-generated takeaways. Using the optimal configurations identified in Experiment 1, we generated 30 chart takeaways across four bar chart layouts using two datasets. We found that LLMs, unlike humans, were generally insensitive to bar layouts, while providing an in-context example helped align model generations in some cases.

Experiment 3: Effect of context and data. If LLMs are not perceptually sensitive to bar layouts, what are they sensitive to? We observed that the comparison types in LLM takeaways varied greatly between charts with the same layout but different contexts and data values. We showed that humans did not exhibit this sensitivity, thus revealing inconsistency as another weakness of LLMs. We further demonstrated that context affected comparison types more than data did for most layouts.

Our work is a first step towards understanding the perceptual awareness of LLMs for visualization tasks. In our case study on bar charts, we not only discovered that LLMs were generally incapable of writing human-aligned takeaways, but they sometimes also generated semantically repetitive and factually incorrect takeaways. We hope our work can motivate investigation into other visualization types and additional dimensions of perceptual awareness. By exposing the weaknesses of LLMs in reading charts, our work outlines challenges in employing LLMs for visualization tasks and informs future LLM development.

2 Related Work

Design choices can influence what patterns people see in data [13, 23, 25, 26, 33]. For example, showing data as a bar chart is more likely to elicit discrete comparisons (e.g., A is larger than B), whereas a line graph facilitates the detection of temporal trends (e.g., X fluctuates as time passes) [3, 29, 40, 44]. Understanding visual affordances can inform visualization tool design [15]. For example, a visualization tool can incorporate affordance mappings as rules to recommend designs that help an analyst see the "right" pattern in the data [45]. These mappings also support the automatic generation of appropriate captions for visualizations [6, 20]. However, the process of identifying visualization affordances across chart types, datasets, design features, and user expertise can be extremely time- and resource-demanding [5, 14, 39]. It is also challenging to incorporate such rules into visualization systems given the diverse possible use cases and syntactic and semantic variations in users' natural language input.

2.1 LLMs for Visualization

LLMs offer exciting possibilities for the visualization community [28, 43]. Not only do they provide a unified natural language interface for a suite of visualization tasks, but they also have the potential to encode visualization knowledge and best practices due to their exposure to Internet-scale training data. A natural application of LLMs for visualization is natural language to visualization (NL2VIS). For instance, Wu et al. [37] explored using LLMs for NL2VIS and achieved state-of-the-art performance. Tian et al. [31] integrated LLMs into a step-by-step reasoning pipeline capable of generating visualizations from abstract natural language.

There is also an increasing body of work that explores using LLMs to generate natural language utterances for visualizations. For example, Tang et al. [30] assembled a dataset of human-written visualization captions and provided initial results on generating captions with a vision-language model. Ko et al. [21] generated natural language datasets from Vega-Lite charts using GPT-4 with some success. In addition, LLMs have been used to generate data presentations [34], enhance visual analytics [46], and guide chart reading [12]. However, to the best of our knowledge, no work so far has assessed the extent to which LLMs are perceptually aware in visualization tasks like humans are. This motivates us to contribute a case study as an initial step toward understanding how well LLMs can predict human takeaways from visualizations.

2.2 Comparison Types in Bar Charts

The spatial layout of data can change which data points people compare and what they take away [14, 39]. For example, as shown in Figure 2, when people look at bar charts in the adjacent layout, they tend to pick out one bar and compare it to multiple other bars. If the bars are in a vertical layout, people would instead compare the vertically aligned bars. Xiong et al. [39] described 12 type of comparisons people engage in when reading bar charts. These categories are grounded and refined based on empirical data [15]. We describe the 12 categories of comparisons below to contextualize the current work. We first introduce the concept of **cardinality**, which has to do with how people group data values prior to making comparisons. The 12 comparison categories can be divided into four cardinalities:

one-to-one (C1,C5, C9): Comparison between one bar and another bar.

two-to-two (C2, C6, C10)): Comparison between one set of two bars and another set of two bars.

all elements (C3, C7, C11): Comparison between one set of bars and the remaining set of bars.

one-to-multiple (C4, C8, C12): Comparison between one bar and a set of bars.

Each cardinality consists of three types of comparison approaches. Referencing the example from Figure 4, we refer to the two companies, A and B, as two "groups", and each of the three data points associated with the store revenues within each group as "elements".

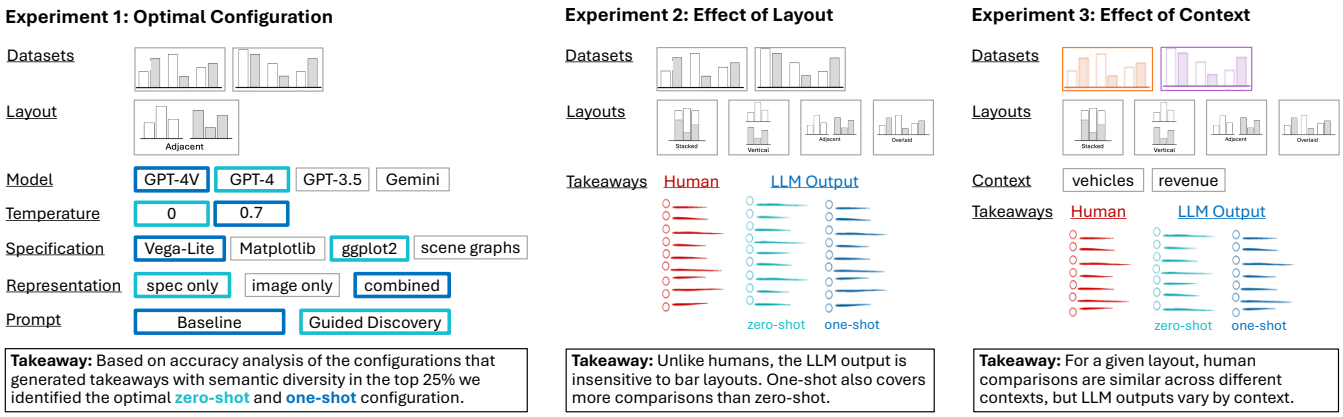

Fig. 3: Our case study includes three experiments. In Experiment 1, we varied the LLM, decoding temperature, chart specification, and prompting strategy, and identified optimal configurations to elicit LLM chart takeaways for both zero-shot and one-shot settings. In Experiment 2, we generated takeaways using optimal configurations and examined whether LLMs' comparisons are perceptually sensitive to bar arrangement like humans are. In Experiment 3, we examined whether LLMs' comparisons are insensitive to data and context like humans are.

Across group - Within element (C1, C2, C3, C4): The reader identifies the same element(s) in each group and compares them.

Across group - Across element (C5, C6, C7, C8): One element from one group is compared to a different element in another group.

Within group - Across element (C9, C10, C11, C12): The reader identifies one group and compares different elements within that group.

In general, amongst the popular comparison types, C1 and C11 are commonly compared for all bar layouts. C3 and C4 are especially associated with the overlaid layout. C9 is commonly made when readers see a vertical bar chart. More detailed distributions of comparison types can be found in Xiong et al. [39].

3 Experiment 1: Optimal Configurations

The appropriateness of chart takeaways generated by LLMs is informed by several parameters, including (1) the choice of which LLM to use (e.g., GPT-3.5, GPT-4, and Llama 2 [32]), (2) the LLM temperature setting, which dictates the randomness in the model's responses, (3) the input chart specification (i.e., how the chart is represented), and (4) the prompting strategy (i.e., how natural language prompts are used to instruct the LLM to generate chart takeaways).

In addition to these four parameters, research has shown that providing LLMs with in-context examples can boost their performance [8]. To this end, we tested LLM performance in two settings: zero-shot, where we requested takeaways without any example chart or takeaways, and one-shot, where we provided the LLMs with a sample chart and human takeaways before requesting takeaways on the test case.

In order to generate human-aligned chart takeaways, we first need to ensure that the generations are accurate. In addition, since humans tend to write diverse takeaways for any given chart, an optimal configuration must be capable of generating semantically diverse takeaways. In Experiment 1, we sought to determine a configuration for the four parameters to optimize both takeaway accuracy and semantic diversity. Using stimuli from Xiong et al. [39], we partially replicated their study by prompting LLMs to produce 30 semantically diverse takeaways and assessed takeaway accuracy and semantic diversity for visualizations.

3.1 Model Types and Temperature Settings

We generated takeaways using four state-of-the-art LLMs at the time of experiment: GPT-4-1106-vision-preview[1] (hereafter GPT-4V), GPT-4-0613[2] (hereafter GPT-4), GPT-3.5-turbo-1106[3] (hereafter GPT-3.5), and Gemini 1.0 Pro[4]. For each model type, we experimented with two temperature settings. The temperature values were set at one of {0, 0.7}. A lower temperature value results in more focused and deterministic output, while a higher value allows for more randomness and creativity in the responses. We chose these two values because a temperature of 0 corresponds to greedy decoding and minimizes randomness, whereas a value of 0.7 is a typical default setting for many LLMs.

3.2 Datasets and Chart Specifications

Xiong et al. [39] generated two datasets for their stimuli, each containing two groups of three data points. They visualized the data using bar charts and tested four spatial layouts—vertically juxtaposed, horizontally juxtaposed, overlaid, and stacked (see Figure 2)—resulting in eight visualizations. Figure 4 shows an example visualization using the adjacent layout to depict the revenues of two companies (Company A and Company B) across three stores (Store 1, Store 2, and Store 3).

To generate chart takeaways using LLMs, we must represent charts in a way that can be recognized by the models. Inspired by Tang et al. [30], we experimented with three classes of chart representations in this work: text-based, image-based, and a combination of both text and image. For text-based representations, we explored four chart representations: Vega-Lite [27], Matplotlib [19], ggplot2 [36], and scene graphs, which are hierarchical representations of the visual elements in a visualization. In order to authentically reproduce Xiong et al. [39], we made sure that these textual representations, when rendered as images, looked similar to the original stimuli that human participants saw. For image-based representations, we simply represented visualizations as bitmap images. In addition, we tried combining one of the text-based representations with images. For instance, in one experiment, we fed both the image (Figure 4) and the Matplotlib specification producing it to GPT-4V. In total, we considered nine chart representations (four text-based, one image-based, and four combined). We include all eight stimuli and their specifications in the supplemental materials.

3.3 Prompting

Prompting is essential to the performance of LLMs (e.g., [42], [35]). Good prompts employ appropriate strategies and articulate the tasks clearly. In this work, we explored two prompting strategies: the baseline strategy and the guided discovery strategy [21] (detailed in Section 3.3.2). To understand what language to use for system prompts (instructions guiding the behavior of the LLM) and user prompts (specific user queries), we piloted several prompts on GPT-4 using the baseline strategy but with slight variations in wording. For instance, we tested system prompts with different levels of specificity, such as "You are a helpful research assistant" and "I am a visualization researcher and you are a helpful research assistant. You should draw on your knowledge of graphical perception research to predict what humans will write as takeaways to visualizations." We adopted the task description and

[1] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
[2] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo
[3] https://platform.openai.com/docs/models/gpt-3-5
[4] https://deepmind.google/technologies/gemini/#gemini-1.0

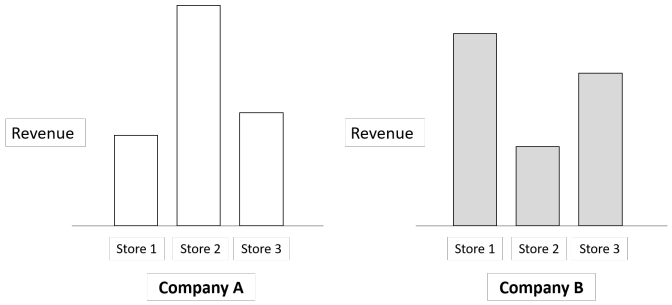

Fig. 4: A horizontally juxtaposed bar chart depicting the revenue of three stores from two companies. Figure is from Xiong et al. [39].

system prompt yielding takeaways with good semantic diversity and high accuracy, and adapted them to the prompting strategies.

3.3.1 Baseline Strategy

This approach begins with a straightforward framing of the task, followed by the chart representation. See a zero-shot example below:

```
System Prompt: I am a visualization researcher and
you are a helpful research assistant. You should draw
on your knowledge of graphical perception research
to predict what humans will write as takeaways to
visualizations.

User Prompt: I will show you some code below,
which generates a chart. The chart depicts the
revenue of three stores selling computers from two
companies. The three stores are Store 1, Store 2,
and Store 3, and the two companies are Company A and
Company B. Your job is to predict what humans will
write as takeaways. Note that humans only see the
visualization. Since you are a language model, I
will show you the code producing the visualization.
Remember, your takeaway should not only be a
description of the chart; it should encapsulate
a take-home message from the chart. Generate 30
semantically different takeaways for this chart. It's
okay for takeaways to not be full sentences.

« Chart specification omitted here for conciseness. »
```

3.3.2 Guided Discovery

The guided discovery strategy was proposed by Ko et al. [21] in their framework generating natural language datasets from Vega-Lite specifications and drew from chain-of-thought prompting [35] and educational psychology [7]. In guided discovery, the user provides scaffolding and poses key questions in the prompt to help the model reason and extract insights. We instructed the model to reason step-by-step: first, we provided the chart and asked the model to extract the chart type and the variables depicted for grounding purposes (scaffolding); then, we instructed the model to attend to bar heights and reason about the pattern shown; finally, we requested 30 semantically distinct takeaways from the visualization. See the following zero-shot example:

```
« System prompt omitted here for conciseness. »

User Prompt: I will show you some code below,
which generates a chart. The chart depicts the
revenue of three stores selling computers from two
companies. The three stores are Store 1, Store 2,
and Store 3, and the two companies are Company A and
Company B. Your job is to predict what humans will
write as takeaways. Note that humans only see the
visualization. Since you are a language model, I will
show you the code producing the visualization.

« Chart specification omitted here for conciseness. »

Let's think step by step. What type of chart is this?
What are the variables depicted?

Assistant: « Response omitted. »

User Prompt: Look at the relative heights of the bars.
Briefly, what patterns do the bars show?

Assistant: « Response omitted. »

User Prompt: Now, generate 30 semantically different
takeaways for this chart. Remember, your takeaway
should not only be a description of the chart; it
should encapsulate a take-home message from the chart.
It's okay for takeaways to not be full sentences.
```

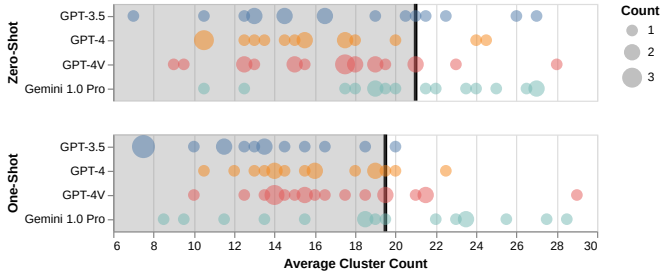

Fig. 5: Distribution of the average cluster count for each configuration (broken down by the LLM type) for the zero-shot and one-shot settings. We reviewed the accuracy of the top 25% (see Section 3.5.1), corresponding to a threshold of 21 for the zero-shot setting and a threshold of 19.5 for the one-shot setting.

3.3.3 Prompts for one-shot settings

The examples in the previous sections are templates for zero-shot settings. For one-shot settings, we provided a sample bar chart using the same spatial layout and specified in the same format as the test case and approximately 150 human takeaways before asking the model for takeaways on the test case. We also tested both the baseline and the guided discovery strategies for one-shot prompts. We largely adhered to the task descriptions and system prompts previously identified. Please refer to the supplemental materials for our one-shot templates.

3.4 Experiment 1: Procedure and Setup

The main goal of Experiment 1 is to identify optimal configurations of the four parameters outlined in Section 3 for zero-shot and one-shot settings. To this end, we prompted LLMs to generate 30 semantically diverse takeaways for each input chart specification. As discussed in the previous sections, we evaluated four LLMs, two temperatures, nine chart representations[5], and two prompting strategies on two datasets in both zero-shot and one-shot settings. Given the large number of configurations, we only experimented on the adjacent layout in Experiment 1 to derive optimal configurations. This resulted in a total of 136 trials across both zero-shot and one-shot settings. To ensure the integrity and independence of our results, we conducted each trial through separate API calls. Finally, we evaluated the takeaways according to the procedure in Section 3.5 and identified the optimal configurations.

3.5 Experiment 1: Evaluation Approaches

We evaluated the takeaways from each configuration on two metrics: semantic diversity and factual accuracy.

[5]Note that we tested chart specifications involving bitmap images only on GPT-4V since it was the only LLM with vision capability among the four tested.

3.5.1 Semantic Diversity

Semantic diversity measures the semantic variety of takeaways, via a "cluster count" proxy. We passed the takeaways for each configuration to GPT-4 and instructed it to cluster them based on their semantics. Take Figure 4 as an example, the takeaway "the revenue of Company A at Store 3 is less than that of Company B" is semantically equivalent to "the revenue of Company B at Store 3 is more than that of Company A". Therefore, we consider these takeaways to be in the same cluster. These takeaways have different meanings from "Store 2 generates the highest revenue for Company A", and thus we consider them to be from different clusters. Hence, for a set of takeaways, the higher the cluster count, the more semantically diverse it is.

Since the prompts request 30 semantically distinct takeaways, we first filtered out configurations that generated semantically repetitive takeaways. Recognizing that in-context examples could influence the generated outcomes, we conducted the filtering process separately for zero-shot and one-shot settings. To verify that the cluster counts returned by GPT-4 were reasonable measures of semantic diversity, we selected 10 configurations at the 10^{th}, 20^{th}, ..., and 100^{th} percentile of cluster count, manually clustered them, and compared machine and human-generated cluster counts.

Next, we created two rankings for all configurations based on semantic diversity as determined by GPT-4, one each for the zero-shot and one-shot settings. We only reviewed configurations generating takeaways in the top 25% in terms of semantic diversity for accuracy of their takeaways and discarded the rest for insufficient semantic diversity.

3.5.2 Factual Accuracy

Factual accuracy measures if takeaways accurately reflect information in the visualization. Two coders manually coded all takeaways for factual accuracy from the top 25% of configurations in terms of semantic diversity. Initially, the two coders separately coded 30 model-generated takeaways for a given visualization. Next, they discussed and established guidelines to categorize the takeaways until reaching a consensus on which ones were accurate, ambiguous, and inaccurate. For example, in Figure 4, "Store 2 generates much higher revenue than Store 1 does for Company A" is an accurate takeaway, "the strategy a company uses affects its revenue" is an ambiguous takeaway, and "Store 1 generates less revenue than does Store 3 for Company B" is an incorrect takeaway.

The two coders each coded half of the takeaways and calculated the percentages of correct, ambiguous, and inaccurate takeaways for each configuration. Since we used each configuration to generate takeaways for two charts, one coder coded each chart and the results from both coders were averaged to obtain the factual accuracy score for that configuration. This approach counterbalances inter-rater differences.

A set of takeaways is considered more factually accurate than another set if it has a higher percentage of correct takeaways. When the percentages of correct takeaways are equal, the set with the lower percentage of factually inaccurate takeaways is deemed more accurate.

3.6 Experiment 1: Results

In this section, we first confirm that GPT-4's clustering results are reasonable proxies for semantic diversity. Then, we present results on the semantic diversity of takeaways produced by each configuration. Finally, we report the optimal configurations for zero-shot and one-shot settings based on factual accuracy.

3.6.1 Confirming GPT-4's Clustering Results

We first verified that the cluster counts generated by GPT-4 were reasonably accurate. After sampling ten sets of takeaways with a diverse range of GPT-4-generated cluster counts following the procedure outlined in Section 3.5.1 and manually clustering them, we performed linear regression to assess the relationship between the GPT-4-generated cluster counts and the manually obtained cluster counts. Results indicated that GPT-4 cluster counts indeed predicted manual cluster counts quite well ($r^2 = 0.83$, $p < 0.001$). Hence, we concluded that the cluster counts generated by GPT-4 were good proxies for semantic diversity.

Table 1: Top five configurations in terms of factual accuracy in the zero-shot setting and their average number of accurate (✓), ambiguous (***?***), and inaccurate (✗) takeaways across the two visualizations tested.

LLM	Temp.	Chart Spec.	Prompt	✓	*?*	✗
GPT-4	0	ggplot2	CoT	23	5	2
GPT-4V	0.7	VL+image	baseline	21.5	6	2.5
GPT-4V	0.7	matplotlib	baseline	21	6.5	2.5
GPT-4V	0	scene+image	CoT	20.5	8	1.5
GPT-4	0	ggplot2	baseline	20	2	8

Table 2: Top five configurations in terms of factual accuracy in the one-shot setting and their average number of accurate (✓), ambiguous (***?***), and inaccurate (✗) takeaways across the two visualizations tested.

LLM	Temp.	Chart Spec.	Prompt	✓	*?*	✗
GPT-4V	0.7	VL+image	baseline	24	4.5	1.5
GPT-4	0	Vega-Lite	CoT	24	1.5	4.5
GPT-4V	0	scene+image	CoT	23.5	2	4.5
GPT-4	0	matplotlib	baseline	22.5	2	5.5
GPT-4	0	scene graph	baseline	22	2	6

3.6.2 Semantic Diversity by Configuration

All configurations tested successfully generated 30 takeaways for each visualization. We visualize the distribution of cluster counts as determined by GPT-4 for every set of zero-shot and one-shot takeaways in Figure 5, broken down by the LLM type. In either case, **there is a wide range of semantic diversity across configurations**.

After retaining the top 25% of configurations as described in Section 3.5.1, we were left with configurations with an average cluster count on the two sets of takeaways of above 21 in the zero-shot setting and those with an average above 19.5 in the one-shot setting. This procedure left us with 19 configurations to code for accuracy in the zero-shot setting and 18 in the one-shot setting. In both settings, Gemini 1.0 Pro produced the highest number of semantically diverse takeaways.

3.6.3 Accuracy by Configuration

Table 1 and Table 2 show the configurations in the top five in terms of average factual accuracy for the zero-shot and one-shot settings, respectively. While there were a mix of temperatures, chart representations, and prompting strategies in the top five, all configurations were notably either generated by GPT-4 or GPT-4V, which suggests **the GPT-4 family's dominance over other LLMs in generating accurate takeaways**. We also found that Gemini 1.0 Pro, while capable of writing semantically diverse takeaways, struggled in factual accuracy. In fact, most configurations involving Gemini 1.0 Pro averaged below 33.3% in accuracy and the majority of takeaways were ambiguous. Also underperforming in accuracy was GPT-3.5, which often produced more factually inaccurate takeaways than accurate ones. Overall, our results show that **generating accurate chart takeaways is still a challenging task for LLMs**.

Following the procedure in Section 3.5, we identified representing charts with ggplot2, utilizing GPT-4 with a temperature of 0, and employing the guided discovery strategy as the optimal configuration for the zero-shot setting and representing charts with Vega-Lite and image, using GPT-4V with a temperature of 0.7, and employing the baseline strategy as the optimal configuration for the one-shot setting.

Due to the black-box nature of LLMs, it is challenging to determine why certain configurations outperformed others. Nonetheless, our results suggest that the choice of LLM has the largest impact on the quality of the generated takeaways. While different temperature values, chart specifications, and prompting strategies appear in the top-performing configurations, the model types remain surprisingly

consistent, underscoring the importance of choosing the right LLM for generating takeaways. This type of qualitative performance improvement resulting from using a superior LLM has been similarly observed across many tasks [9].

4 Experiment 2: Alignment of Human & LLM Takeaways

In Experiment 1, we generated takeaways from adjacent bar charts and identified optimal configurations for zero-shot and one-shot settings. Next, we started to more generally understand whether LLMs are influenced by the spatial layouts of bars like humans are. We tested all four layouts in Figure 2, using the optimal configurations identified in Experiment 1. We prompted LLMs to generate 30 takeaways for each visualization as if it were 30 people in both zero-shot and one-shot settings. In addition to coding takeaways for accuracy, we classified them along the taxonomy Xiong et al. developed, which captures 12 types of comparison takeaways can make (denoted C1, C2, ..., C12, see Section 2.2). By comparing the distributions of LLM-generated takeaways vs. human-written takeaways along these comparison types, we can understand to what extent takeaways from LLMs emulate those from humans.

4.1 Experiment 2: Procedure and Setup

We used the same two datasets as in Experiment 1. We started by replicating all eight original stimuli across the four bar chart layouts from Xiong et al. in ggplot2 and Vega-Lite, which were optimal chart specifications for zero-shot and one-shot settings identified in Experiment 1. We then generated takeaways for every stimulus. In the zero-shot setting, we represented charts using ggplot2, adopted GPT-4 as the LLM, set its temperature to 0, and prompted the model following the guided discovery strategy. In the one-shot setting, we represented charts using both Vega-Lite specifications and bitmap images, adopted GPT-4V as the LLM, set its temperature to 0.7, and prompted the model following the baseline strategy. The prompts we used in Experiment 2 were exactly the same as the ones in Experiment 1, except that instead of asking the model to "generate 30 semantically different takeaways for this chart", we asked it to "generate 30 takeaways for this chart as if you were 30 people".

4.2 Experiment 2: Evaluation Approaches

We evaluated the takeaways generated by LLMs across two dimensions: factual accuracy and human-LLM alignment of comparison types.

Factual accuracy was measured in the same way as in Experiment 1. One coder coded all charts for factual accuracy. We also leveraged the taxonomy from Xiong et al. [39], and one coder coded all the takeaways for comparison types. Since we observed that some LLM-generated takeaways did not make any comparisons, we added an additional category, “Other”, to capture them. For each bar chart layout, we identified the top three most common comparison types averaged across the two examples in model-generated takeaways (denoted as the set M) and in human-written takeaways (denoted as the set H). We calculated the overlap between M and H, a metric we will shorthand as precision@3:

$$\text{precision@3} = \frac{|H \cap M|}{3}.$$

To quantify the distances between the distributions of LLM and human comparison types, we first normalized the frequency distributions into probability distributions (P_M for model and P_H for human) by dividing the number of takeaways falling into each comparison type by the total number of takeaways. Next, we calculated Total Variation Distance (TVD) between P_M and P_H, which is half of the sum of the absolute differences between the probabilities for each of the thirteen categories (the original twelve categories from Xiong et al. [39] and the “Other” category). The value of TVD ranges between zero and one. The closer TVD is to zero, the closer the two distributions are. More formally, TVD between P_M and P_H is defined as:

$$\text{TVD}(P_H, P_M) = \frac{1}{2}\sum_{i=1}^{13} |p_{M_i} - p_{H_i}|,$$

Table 3: Zero-shot takeaway accuracies for the four layouts. Even state-of-the-art LLMs can struggle to write factually accurate takeaways.

Layout	Accurate	Ambiguous	Inaccurate
Adjacent	68.33%	15.00%	16.67%
Overlaid	80.00%	15.00%	5.00%
Stacked	21.67%	18.33%	60.00%
Vertical	56.67%	3.33%	40.00%

Table 4: One-shot takeaway accuracies for the four layouts. One-shot accuracies improve upon zero-shot accuracies in most cases.

Layout	Accurate	Ambiguous	Inaccurate
Adjacent	70.00%	3.33%	26.67%
Overlaid	65.00%	5.00%	30.00%
Stacked	51.67%	3.33%	45.00%
Vertical	73.33%	0.00%	26.67%

Table 5: TVD’s between the comparison type distributions of zero-shot takeaways and human takeaways, and between one-shot takeaways and human takeaways for each layout. One-shot TVDs are lower in most cases, suggesting the benefit of providing human-written takeaways in calibrating LLM generations.

Layout	Zero-Shot TVD	One-Shot TVD
Adjacent	0.3456	0.2090
Overlaid	0.5278	0.3778
Stacked	0.3699	0.2903
Vertical	0.2167	0.3679

4.3 Experiment 2: Results

In this section, we report how the optimal configurations performed on takeaway accuracy and alignment with human comparison types.

4.3.1 Accuracy

Tables 3 and 4 show the accuracies of takeaways across zero-shot and one-shot settings, respectively. In the zero-shot setting, accuracies varied significantly across layouts, with a minimum of 21.67% for the stacked layout and a maximum of 80.00% for the overlaid layout. We observed that **in many cases LLMs suffered from significant hallucination**. For instance, when generating zero-shot takeaways for the stacked layout, 60% of takeaways were factually inaccurate. Upon examining these incorrect takeaways, we found that the LLM consistently confused the group labels. This points to weaknesses even in state-of-the-art LLMs to accurately read and reason about visualizations.

The one-shot setting yielded more consistent accuracies across different layouts in the range of 51.67% to 73.33%. We also found that one-shot accuracies were higher than zero-shot accuracies in three of the four layouts, suggesting **the benefits of in-context learning on grounding takeaway generation.** In particular, the presence of example takeaways boosted the accuracy for the stacked layout by 30%. Furthermore, one-shot takeaways tended to contain fewer ambiguous statements. That said, at least 26.67% of one-shot takeaways in every layout did not accurately portray the visualization. Hence, there is still much room for improvement when it comes to generating accurate chart takeaways with LLMs.

4.3.2 Alignment of Comparison Types

Figure 6 depicts the distributions of comparison types of human-written takeaways, LLM zero-shot takeaways, and LLM one-shot takeaways for the four layouts. Visually, the distributions of zero-shot takeaways tend to deviate from the human takeaways (e.g., zero-shot takeaways place an overly very high probability mass on C12 for the adjacent layout), while the distributions of one-shot takeaways look more similar

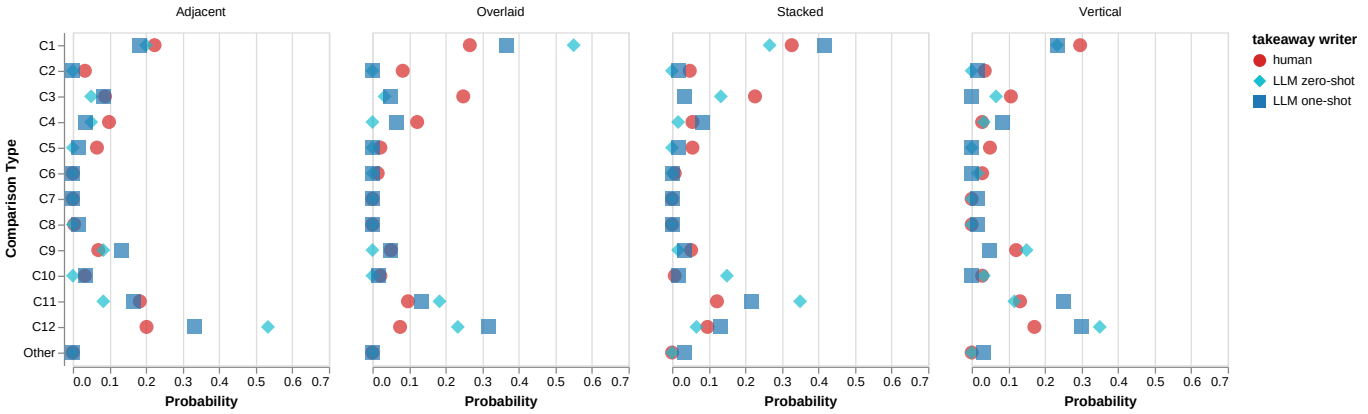

Fig. 6: Distributions of human, LLM zero-shot, and LLM one-shot comparison types for each layout. LLM one-shot distributions are generally closer to human ones than zero-shot distributions.

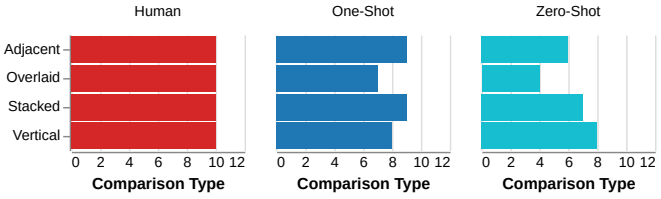

Fig. 7: The number of unique comparison types (excluding the "Other" category) generated by humans, LLM one-shot, and LLM zero-shot settings. The one-shot setting generates more unique comparison types than the zero-shot setting.

to human-written takeaways. Table 5 shows the TVD's between the distributions of comparison types for LLM zero-shot takeaways and human takeaways, and between those for LLM one-shot takeaways and human takeaways. In all but the vertical layout, TVD is lower in the one-shot setting, confirming the visual conclusion. Even though our prompt did not outline alignment in comparison types as an optimization objective, **the LLM better calibrated the comparison types in its generations when an example was provided**.

Figure 7 shows the number of unique comparison types (excluding the "Other" category) in human, LLM zero-shot, and LLM one-shot takeaways. While human takeaways consistently covered a diverse set of comparison types (*mean* = 10), the LLMs tended to focus on fewer comparison types across all layouts in the zero-shot setting (*mean* = 6.25). When provided with example human takeaways, the LLMs showed expanded coverage of comparison types (*mean* = 8.25), demonstrating the benefits of in-context learning for simulating the diverse takeaways humans tend to write.

Table 6 further details the top three most frequent comparison types for each layout from each source while Table 7 shows precision@3. Compared to zero-shot takeaways, one-shot takeaways attained higher precision@3 for two layouts and equal precision@3 for the remaining two, indicating that **the presence of example takeaways tended to allow LLMs to better capture the most afforded comparison types**. While LLMs were generally good at predicting frequent comparison types across all layouts, such as C1, C11, and C12, when rarer comparison types made it to the top three, such as C3 and C4, LLMs failed to predict them. However, the frequent appearance of a less common comparison type in other layouts within the top ranks for a specific layout suggests a unique affordance by that layout for such comparisons. Failure by the LLM to capture these affordances is further evidence that **LLM comparison types are generally insensitive to bar layout**.

5 Experiment 3: Effect of Context and Data

In Experiment 2, we calculated the average distributions of LLM comparison types across two examples for each layout, and compared them

Table 6: The top three most frequent comparison types generated for each layout in the human, LLM zero-shot, and LLM one-shot takeaways. LLMs are good at predicting frequent comparison types across all layouts, but fail to predict when generally rarer types are in the top three.

Layout	Human			Zero-Shot			One-Shot		
Adjacent	C1	C12	C11	C12	C1	C9=C11	C12	C1	C11
Overlaid	C1	C3	C4	C1	C12	C11	C1	C12	C11
Stacked	C1	C3	C11	C11	C1	C10	C1	C11	C12
Vertical	C1	C12	C11	C12	C1	C9	C12	C11	C1

Table 7: Precision@3 for predicting the top comparison types from human takeaways. One-shot precisions are at least as high as zero-shot precisions in all cases. In the case where two comparison types tied for third place, each was counted as 0.5 comparison types.

Layout	Zero-Shot	One-Shot
Adjacent	83.33%	100.00%
Overlaid	33.33%	33.33%
Stacked	66.67%	66.67%
Vertical	66.67%	100.00%

to distributions derived from human takeaways. The results suggested that LLMs are generally *not* sensitive to spatial layouts like humans are. However, before aggregating the distributions of LLM comparison types, **we frequently observed significant variations across the two examples within the same layout**. For instance, in the stacked layout, while half of the takeaways involved across-group within-element one-to-one comparisons (C1) for one example chart, only one out of 30 takeaways made this type of comparison for the other example (see Figure 8). The two example charts for each layout depicted different contexts (e.g., prices of three drinks in two bars or popularity of three bands in two countries) and different data values. It is possible that the LLMs overly relied on the context and data to generate takeaways compared to humans. We ask two questions in this experiment. First, are humans and LLM similarly sensitive to context and data when generating takeaways? Second, if not, what drives the instability in comparison types across examples in the same layout?

5.1 Experiment 3: Procedure, Setup, and Evaluation

We focused our analysis on human takeaways and LLM zero-shot takeaways as a case study. To answer the first question, we visualized and compared both human and LLM comparison type distributions for the two examples for each layout. For both the human the LLM distributions, we calculated and compared the TVD between the comparison

Table 8: TVD's between the comparison type distributions for the two example charts in human takeaways and LLM takeaways for each layout. LLM shows much more variation than humans do.

Layout	Human TVD	LLM TVD
Adjacent	0.2120	0.4333
Overlaid	0.2661	0.4333
Stacked	0.0998	0.6333
Vertical	0.3542	0.5000

distributions of the two examples. We also compared Spearman's rho rank correlation between the frequency rankings of each comparison type across the two examples for both human and LLM distributions.

To answer the second question, we generated eight new bar charts by modifying the original eight stimuli from Xiong et al. For each chart, we preserved its context while introducing a new dataset. For example, if the existing two stacked bar charts visualized "context 1 and dataset 1" and "context 2 and dataset 2", we created two new stacked bar charts visualizing "context 1 and dataset 2" and "context 2 and dataset 1". Next, we generated zero-shot takeaways using the optimal configuration established in Experiment 1, and the same coder from Experiment 2 coded these takeaways for comparison types. We then visualized the LLM comparison type distributions for the four example charts for each layout. To examine the effect of contexts on distributions, we computed the TVD's between the distributions for each pair of charts showing the same dataset but different contexts and calculated the average. To examine the effect of data on distributions, we computed the TVD's between the distributions for each pair of charts showing the same context but different datasets and calculated the average.

5.2 Experiment 3: Results

We first report on the stability of human comparison type distributions across examples within the same layout. We then present evidence that variations in LLM comparison type within the same layout is more attributable to context than to data in most cases.

5.2.1 Human comparison type distributions are stable across examples within the same layout

Figure 9 visualizes the human and LLM comparison type distributions for the two example charts from Xiong et al. for each layout. While there is much variation in LLM comparison type distributions across the two examples, human distributions does not appear very different. Table 9 shows the TVD's between the comparison type distributions for the two example charts in human takeaways and LLM zero-shot takeaways for each layout. In every layout, LLM TVD is larger than human TVD. Table 8 also suggests that while the frequency rankings of human comparison types are all strongly correlated between examples in each layout, those of LLM comparison types are only moderately strongly correlated in two layouts and uncorrelated in the rest. These results confirm that **the instability in LLM comparison type distributions between examples but within the same layout is not found in human takeaways**. Since the decoding temperature of the GPT-4 was set at 0 when generating these takeaways, this instability was not due to stochasticity in decoding. Hence, we conclude that as far as zero-shot generation is concerned, LLMs are neither accurate nor consistent in writing chart takeaways with human-aligned comparison types.

5.2.2 Context affects LLM comparison types more than data

Figure 9 shows four distributions of comparison types induced by the four charts for each layout. Except for the stacked layout, the distributions visually appear more similar across datasets (same shape, different colors) than across contexts (same color, different shapes). Table 10 confirms that the average TVD between distributions from different contexts is larger than that between distributions from different datasets with the exception of the stacked layout. Therefore, we conclude that **variations in LLM comparison types within the same layout is more attributable to context than to data in most cases**.

Table 9: Spearman's rho correlation of comparison type rankings in humans and LLM across the two examples for each layout. Significance levels: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. While the rankings of comparison types remain stable across different charts with the same layout in human takeaways, they tend to vary greatly in LLM takeaways.

Layout	Spearman's rho (Human)	Spearman's rho (LLM)
Adjacent	0.8937***	0.5127
Overlaid	0.8852***	0.7042**
Stacked	0.9285***	0.3049
Vertical	0.9078**	0.6072*

Table 10: Average TVD's in LLM comparison type distributions between the two pairs of charts showing the same dataset but different contexts, and the two pairs showing same context but different datasets for each layout. Differing contexts tend to induce greater shifts in LLM comparison type distributions.

Layout	Different Datasets	Different Contexts
Adjacent	0.2167	0.3333
Overlaid	0.2333	0.5000
Stacked	0.6333	0.4667
Vertical	0.2167	0.4833

6 Limitations and Future Work

To conclude, we detail several limitations from our investigation that lay the foundation for future research.

Other Chart Types: In this work, we used bar charts with varying layouts as a case study. While we provide a rich set of human takeaways for each chart, future work can expand to cover more chart types (e.g., scatterplots, line charts, maps) and design manipulations (e.g., color choices, presence of embellishments) to test the generalizability of our results and the capability thresholds of LLM predictions. Since our case study requires much manual coding, we also encourage future work to explore scalable ways to examine LLM perceptual awareness. One promising direction is to employ multiple LLMs to perform qualitative coding and cross-validate each other [16]. Another direction is to develop automatic pipelines to elicit and model LLMs' visualization-related behavior.

More Datasets: The stimuli in the present study were created using two datasets and eight contexts. To examine the generalizability of our findings, we recommend future research to test a larger number of datasets with additional contexts. Despite showing poor perceptual alignment on open-domain datasets like the ones we tested, LLMs might produce more human-aligned takeaways on domain-specific datasets. While the pretraining data on open-domain topics include a wide variety of linguistic styles and perspectives, data for certain domains (e.g., protein analysis in scientific visualization [18]) may feature more formulaic language and stronger patterns. This specificity could make it easier for LLMs to generate takeaways that align closely with human analysis. Further, we encourage future work to experiment with datasets of varying complexities. Given a more complex dataset, human chart reading behavior tends to diverge as people attend to different aspects of the dataset and visualization, such as locally noticing a statistic or globally noticing a trend [3, 4, 22]. Thus, employing more complex datasets could similarly offer opportunities to characterize LLM visualization perception on multiple levels and dimensions.

Task-Specific Alignment: Existing work has demonstrated that user takeaways from visualizations can be task-dependent [20, 24]. In the present study, we focused on eliciting overall takeaways from charts. Future work can additionally test whether the misalignment between LLM outputs and human takeaways also takes place when the visualization is provided in the context of a more specific task, such as low-level analytic tasks [2]. Specifically, because humans tend to be prone to cognitive biases even when completing objective analytic tasks with visualizations, such as estimating correlations [17, 41], it would be in-

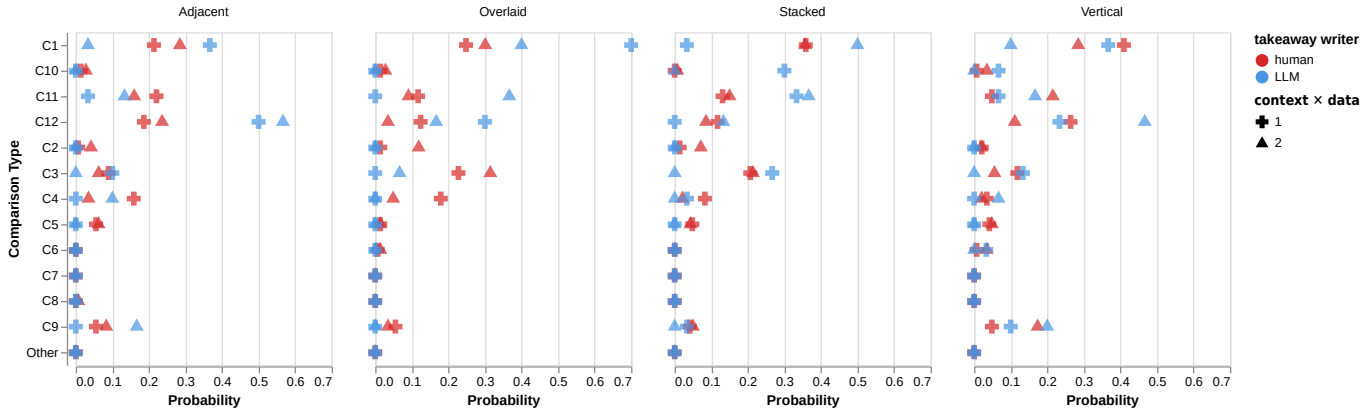

Fig. 8: Distributions of human and LLM comparison types for the two example charts for each layout. Unlike humans, whose distributions remain relatively stable across charts using the same layout but different contexts and data, LLM distributions show significant differences.

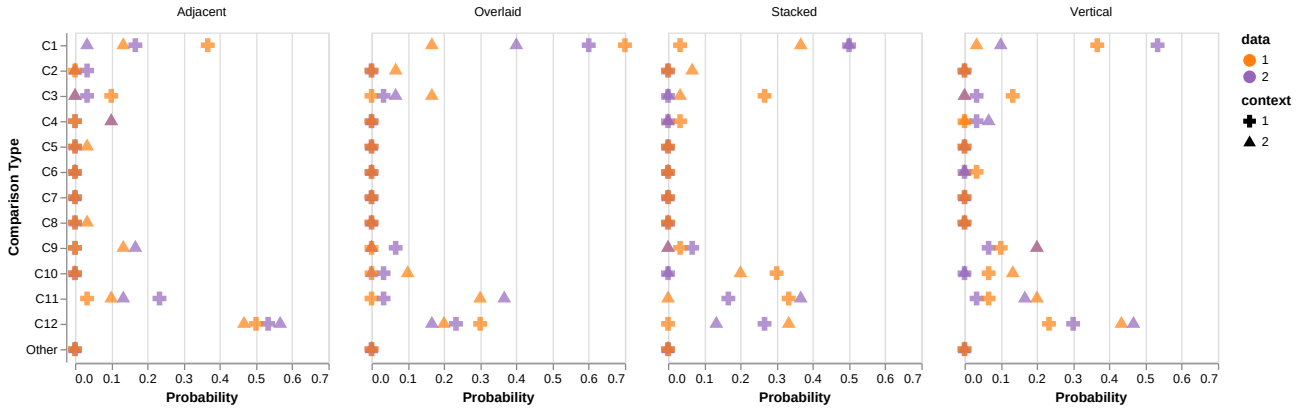

Fig. 9: Distributions of LLM comparison types for the four examples (2 contexts × 2 datasets) for each layout. With the exception of the stacked layout, differing contexts (same color, different shapes) induce greater variations in the distributions than differing data (same shape, different colors).

teresting to explore the extent to which LLMs are capable of predicting these biases in human behaviors.

Prompting and Finetuning: In Experiment 1, we tested for an optimal configuration of chart specifications, temperature, prompting strategies, and LLM type using the adjacent bar layout. Future work can test for optimal configurations across a wider range of charts, layouts, and datasets with an eye towards better generalizability.

Additionally, the two prompting strategies we tested serve as a starting point to demonstrate the power of prompting. Considering the black-box nature of LLMs, there could be an infinite number of ways to prompt LLMs to obtain even stronger results. Future work could explore the design space of prompting LLMs for human-aligned chart takeaways. We suspect adding elements that highlight aspects of the visualization an LLM should pay special attention to, such as pointing out a design feature that has been manipulated, would increase LLM-human takeaway alignment. Future exploration of the prompt space can also consider manipulating the number of examples provided to the LLMs. While the zero-shot and one-shot settings tested in this work are fast and practical, increasing the number of in-context examples has the potential to better calibrate chart takeaways generated by LLMs in light of the benefits of scaling in-context learning in other domains [1]. Beyond prompting, another exciting opportunity for takeaway alignment is finetuning. By finetuning LLMs on pairs of charts and human-written takeaways, we might be able to improve their ability to generate human-aligned takeaways.

General-Purpose Visualization Assistants: Visualization design choices can profoundly impact what people see and take away, making it challenging even for human experts to predict affordances and create effective visualizations [39]. With a deeper understanding of the limitations of LLMs for modeling human perception of visualizations, future work can devise strategies to address these weaknesses and create perceptually-aware, general-purpose visualization assistants. Given sufficient knowledge of visualization affordances, LLMs have the potential to generate human-aligned chart takeaways and assist with various other visualization tasks, such as critiquing existing designs, recommending new designs, and annotating visualizations for better comprehension. Therefore, if we are able to provide appropriate guardrails and develop effective human-LLM collaboration frameworks, LLMs could be able to democratize visualization tasks for laypeople.

7 Conclusion

Through a case study on bar charts with different layouts, we provide an initial attempt at understanding the extent to which LLMs are perceptually aware when generating visualization takeaways. In Experiment 1, we identified the optimal configurations to generate chart takeaways. We found that even state-of-the-art LLMs can sometimes struggle to generate semantically diverse and factually accurate takeaways. In Experiment 2, we prompted LLMs to generate human-aligned takeaways across multiple spatial layouts. We found that LLMs are generally not sensitive to bar chart layout like humans are. Nonetheless, providing an example chart and some human takeaways helps align model generations in some cases. In Experiment 3, we observed that, across different data and contexts, while humans generally reported similar comparison types across charts with the same layout, LLMs showed tremendous variability in what they compared. This shows that LLMs are too reliant on context and data when generating takeaways. We further demonstrated that context affects LLM comparison types more than data for most layouts.

Acknowledgments

We thank our reviewers for their helpful feedback. This work was partially supported by NSF award IIS-2237585 and IIS-2311575.

References

[1] R. Agarwal, A. Singh, L. M. Zhang, B. Bohnet, S. Chan, A. Anand, Z. Abbas, A. Nova, J. D. Co-Reyes, E. Chu, et al. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*, 2024. 9

[2] R. Amar and J. Stasko. A knowledge task-based framework for design and evaluation of information visualizations. In *IEEE Symposium on Information Visualization*, pp. 143–150, 2004. 8

[3] C. X. Bearfield, C. Stokes, A. Lovett, and S. Franconeri. What does the chart say? grouping cues guide viewer comparisons and conclusions in bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1, 2, 8

[4] C. X. Bearfield, L. van Weelden, A. Waytz, and S. Franconeri. Same data, diverging perspectives: The power of visualizations to elicit competing interpretations. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 8

[5] J. Boy, L. Eveillard, F. Detienne, and J.-D. Fekete. Suggested interactivity: Seeking perceived affordances for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):639–648, 2015. 2

[6] D. Bromley and V. Setlur. What is the difference between a mountain and a molehill? quantifying semantic labeling of visual features in line charts. In *2023 IEEE Visualization and Visual Analytics (VIS)*, pp. 161–165. IEEE, 2023. 2

[7] A. L. Brown and J. C. Campione. *Guided discovery in a community of learners.* The MIT Press, 1994. 4

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 3

[9] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 1, 6

[10] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. d. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021. 1

[11] Y. Chen, R. Li, A. Mac, T. Xie, T. Yu, and E. Wu. Nl2interface: Interactive visualization interface generation from natural language queries. *arXiv preprint arXiv:2209.08834*, 2022. 1

[12] K. Choe, C. Lee, S. Lee, J. Song, A. Cho, N. W. Kim, and J. Seo. Enhancing data literacy on-demand: Llms as guides for novices in chart interpretation. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–17, 2024. doi: 10.1109/TVCG.2024.3413195 2

[13] S. Few. *Information dashboard design: The effective visual communication of data*. O'Reilly Media, Inc., 2006. 2

[14] R. Fygenson, S. Franconeri, and E. Bertini. The arrangement of marks impacts afforded messages: Ordering, partitioning, spacing, and coloring in bar charts. *arXiv preprint arXiv:2308.13321*, 2023. 1, 2

[15] A. Gaba, V. Setlur, A. Srinivasan, J. Hoffswell, and C. Xiong. Comparison conundrum and the chamber of visualizations: An exploration of how language influences visual design. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):1211–1221, 2022. 1, 2

[16] M. Grunde-McLaughlin, M. S. Lam, R. Krishna, D. S. Weld, and J. Heer. Designing llm chains by adapting techniques from crowdsourcing workflows. *ArXiv*, abs/2312.11681, 2023. 8

[17] L. Harrison, F. Yang, S. Franconeri, and R. Chang. Ranking visualizations of correlation using weber's law. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1943–1952, 2014. 8

[18] E. Ho, R. Webber, and M. R. Wilkins. Interactive three-dimensional visualization and contextual analysis of protein interaction networks. *Journal of proteome research*, 7(01):104–112, 2008. 8

[19] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95, 2007. 3

[20] D. H. Kim, V. Setlur, and M. Agrawala. Towards understanding how readers integrate charts and captions: A case study with line charts. In *Proc. of the 2021 CHI Conf. on Human Factors in Comput. Sys.*, article no. 610, 11 pages, pp. 1–11. ACM, New York, NY, USA, 2021. 2, 8

[21] H.-K. Ko, H. Jeon, G. Park, D. H. Kim, N. W. Kim, J. Kim, and J. Seo. Natural language dataset generation framework for visualizations powered by large language models. *arXiv preprint arXiv:2309.10245*, 2023. 2, 3, 4

[22] A. Lundgard and A. Satyanarayan. Accessible visualization via natural language descriptions: A four-level model of semantic content. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):1073–1083, 2021. 8

[23] J. Mackinlay. Automating the design of graphical presentations of relational information. *Acm Transactions On Graphics (Tog)*, 5(2):110–141, 1986. 2

[24] S. Malpica, D. Martin, A. Serrano, D. Gutierrez, and B. Masia. Task-dependent visual behavior in immersive environments: A comparative study of free exploration, memory and visual search. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 8

[25] C. Nothelfer and S. Franconeri. Measures of the benefit of direct encoding of data deltas for data pair relation perception. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):311–320, 2019. 2

[26] B. Saket, A. Endert, and C. Demiralp. Task-based effectiveness of basic visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 2018. 2

[27] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2016. 3

[28] V. Schetinger, S. Di Bartolomeo, M. El-Assady, A. McNutt, M. Miller, J. P. A. Passos, and J. L. Adams. Doom or deliciousness: Challenges and opportunities for visualization in the age of generative models. In *Computer Graphics Forum*, vol. 42, pp. 423–435. Wiley Online Library, 2023. 2

[29] P. Shah and E. G. Freedman. Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in cognitive science*, 3(3):560–578, 2011. 2

[30] B. J. Tang, A. Boggust, and A. Satyanarayan. Vistext: A benchmark for semantically rich chart captioning. *arXiv preprint arXiv:2307.05356*, 2023. 2, 3

[31] Y. Tian, W. Cui, D. Deng, X. Yi, Y. Yang, H. Zhang, and Y. Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2

[32] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 3

[33] B. Tversky. Visualizing thought. In *Handbook of human centric visualization*, pp. 3–40. Springer, 2014. 2

[34] F. Wang, Y. Lin, L. Yang, H. Li, M. Gu, M. Zhu, and H. Qu. Outlinespark: Igniting ai-powered presentation slides creation from computational notebooks through outlines. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024. generates slides from notebooks given outline. 2

[35] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 3, 4

[36] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. 3

[37] Y. Wu, Y. Wan, H. Zhang, Y. Sui, W. Wei, W. Zhao, G. Xu, and H. Jin. Automated data visualization from natural language via large language models: An exploratory study. *Proceedings of the ACM on Management of Data*, 2(3):1–28, 2024. 2

[38] C. Xiong, E. Lee-Robbins, I. Zhang, A. Gaba, and S. Franconeri. Reasoning affordances with tables and bar charts. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 1

[39] C. Xiong, V. Setlur, B. Bach, K. Lin, E. Koh, and S. Franconeri. Visual arrangements of bar charts influence comparisons in viewer takeaways. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 1, 2, 3, 4, 6, 9

[40] C. Xiong, J. Shapiro, J. Hullman, and S. Franconeri. Illusion of causality in visualized data. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):853–862, 2019. 1, 2

[41] C. Xiong, C. Stokes, Y.-S. Kim, and S. Franconeri. Seeing what you believe or believing what you see? belief biases correlation estimation. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 8

[42] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan. Tree of thoughts: Deliberate problem solving with large

language models. *arXiv preprint arXiv:2305.10601*, 2023. 3

[43] Y. Ye, J. Hao, Y. Hou, Z. Wang, S. Xiao, Y. Luo, and W. Zeng. Generative ai for visualization: State of the art and future directions. *Visual Informatics*, 2024. 2

[44] J. Zacks and B. Tversky. Bars and lines: A study of graphic communication. *Memory & cognition*, 27:1073–1079, 1999. 1, 2

[45] Z. Zeng, J. Yang, D. Moritz, J. Heer, and L. Battle. Too many cooks: Exploring how graphical perception studies influence visualization recommendations in draco. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 2

[46] Y. Zhao, Y. Zhang, Y. Zhang, X. Zhao, J. Wang, Z. Shao, C. Turkay, and S. Chen. Leva: Using large language models to enhance visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 2024. 2